# Vehicle Price Analysis and Prediction

## ****Overview****

This project explores the factors driving the price of used cars using a dataset of 426,000 records sourced from Kaggle. The goal is to identify critical attributes influencing car prices and provide actionable recommendations for used car dealerships to optimize their inventory and pricing strategies.

**Data Sources**

The dataset (vehicles.csv) includes information about vehicles, such as:

* **Categorical features**: Vehicle types, make, model, and VIN.
* **Numerical features**: Price, mileage, engine size, and year.
* **Missing data**: Addressed through statistical imputation or removal of irrelevant columns.

**Business Problem**

The primary goal of this project is to address the following questions:

1. **Analysis Goals**:
   * What are the key factors influencing vehicle prices?
   * How do missing values and data inconsistencies affect the dataset?
2. **Prediction Goals**:
   * Develop a reliable model to predict vehicle prices based on relevant features.
   * Provide actionable insights for pricing strategies and market analysis.

## ****CRISP-DM Framework****

The analysis follows the **CRISP-DM** methodology:

1. **Business Understanding**: Understand what makes a car more or less expensive.
2. **Data Understanding**: Explore the dataset and uncover trends and issues.
3. **Data Preparation**: Clean the data, handle missing values, and engineer relevant features.
4. **Modeling**: Use multiple regression models and hyperparameter tuning to predict prices.
5. **Evaluation**: Assess model performance using metrics such as R² and RMSE.
6. **Deployment**: Provide insights and recommendations to aid dealership decision-making.

## ****Key Insights****

### **1. Data Understanding**

* **Dataset**: 426K cars with features such as price, mileage, year, make, model, engine size, and more.
* **Observations**:
  + High variability in car prices, with some extreme outliers.
  + Mileage and year strongly correlate with price.
  + Several features have missing or redundant data.

### **2. Data Cleaning and Preparation**

* Removed duplicate entries based on unique identifiers.
* Handled missing values:
  + Imputed missing mileage and engine size with median values.
* Engineered new features:
  + **An indicator for old or vintage cars.**
  + One-hot encoding for categorical variables (e.g., make and model).
* Final dataset prepared with standardized numerical features for modeling.

### **3. Visualizations**

* **Price Distributions**:
  + Prices are right-skewed, indicating most vehicles are in the lower price range.
* **Correlations**:
  + Strong positive correlation between year (or vehicle age) and price.
  + Mileage negatively correlates with price, indicating depreciation with usage.
* **Density curves and histograms**:
  + Highlight variability in prices across vehicle types, the prices show extreme skewness to the right and many high outliers.

### **4. Modeling**

* **Models Used**:
  + Linear Regression, Quadratic Regression, Ridge Regression, and Lasso Regression.
* **Best Model**: Lasso regression with polynomial features (degree 2) and strong L1 regularization (alpha=100).
* **Hyperparameter Tuning**:
  + Used GridSearchCV with cross-validation to optimize model parameters.

**Interpretation of the GridSearchCV Output:**

* The best model is a quadratic regression with interactions and Lasso regularization.
* The data likely follows a near-linear trend, and higher-degree polynomial terms did not improve performance.
* Strong regularization (alpha=100) implies the dataset might have noisy features or multicollinearity.
* The best model has degree=2 with interaction terms
* **Model and Evaluation**
* **Model**:
  + A Ridge regression model was selected for its robustness to multicollinearity and scalability.
  + Hyperparameter tuning using GridSearchCV improved model performance.
* **Metrics**:
  + **Root Mean Squared Error (RMSE)**: Measures prediction error; lower is better.
  + **Mean Absolute Error (MAE)**: Represents average prediction deviation from actual prices.
  + **R² Score**: Indicates the proportion of variance explained by the model.
* **Results**: The best model is lasso regression with degree=2.
  + RMSE: $7,000
  + R² Score: ~0.70 (indicating that the model explains 85% of price variability).

### **5. Findings**

* **Key Drivers of Price**:
  + Vehicle age (newer vehicles have higher prices).
  + Mileage (lower mileage cars are priced higher).
  + Engine size and make/model contribute significantly to price predictions.
* **Eliminated Redundant Features**:
  + Lasso regularization dropped features with negligible contributions.

### **6. Recommendations**

* **For Inventory Management**:
  + Focus on acquiring newer vehicles with lower mileage, as consumers value these more.
  + Highlight features like engine size and specific makes/models in marketing campaigns.
* **For Pricing Strategies**:
  + Use the predictive model to price vehicles based on mileage, age, and other vital attributes.
  + Adjust pricing strategies based on region-specific demand and feature preferences.
  + The “best” fitted model provides a reliable mechanism for predicting vehicle prices.
  + Understanding mileage, vehicle age, and categorical attributes can guide pricing decisions.
  + Cleaning and encoding categorical features significantly enhanced model accuracy.
* My models focused on cars below the 90th percentile for the price. Cars in the top 10th percentile tier based on descriptive statistics have the following attributes:

**region:** The most frequent was **Seattle-Tacoma**

**price:** Range from 68,000- 200,142 dollars.

**Year:** varied from 1932 to 2021.0, with a median of 2019.

**Manufacturer:** Ford accounted for 352 of the 1178 cars, making it the most popular manufacturer. The Corvette manufactures accounted for 41 of the 1178 cars, making it the most popular model.

**Condition:** excellent accounted for 231 of the 1178 cars

**Cylinders:** 8 cylinders accounted for 426

**Fuel:** diesel accounted for 532

**Odometer:** Ranged from 0 (brand new cars) to 640,626 with a mean of roughly 26,000.

**The State of** California has the most expensive cars, and diesel is the most popular fuel among those.

A mileage of around 26,000 is the median mileage for the top 10% valuable cars.